Title: Comprehensive Analysis and Classification of Wheat Seed Dataset Using Various Machine Learning Techniques

Abstract: This project aims to analyze and classify the wheat seed dataset using a variety of machine learning techniques, including K-Means++, soft K-means, Principal Component Analysis (PCA), Nonlinear Autoencoders, Multi-Layer Perceptron (MLP), Support Vector Machines (SVM), SVM with Gaussian kernel, and AdaBoost. The project will compare the performance of these methods on both multi-class and binary classification tasks.

Dataset Description: The wheat seed dataset is given in seeds_dataset.txt. This dataset contains 210 data samples. Each sample has seven input features and one output label, which is described in the following table.

列号	列名	含义	特征 / 类标记	可取值
1	area	区域	特征	实数
2	perimeter	周长	特征	实数
3	compactness	紧密度	特征	实数
4	length of kernel	籽粒长度	特征	实数
5	width of kernel	籽粒宽度	特征	实数
6	asymmetry coefficient	不对称系数	特征	实数
7	length of kernel groove	籽粒腹沟长度	特征	实数
8	class	类别	类标记	1,2,3

Methodology:

1. **K-Means++ Algorithm:**
   - ❖ Implement the K-Means++ algorithm using Python and NumPy.
   - ❖ Apply K-Means++ to the dataset with K=3 to perform clustering.
2. **Soft K-Means Algorithm:**
   - ❖ Implement the soft K-means algorithm using Python and NumPy.
   - ❖ Apply soft K-means to the dataset with K=3 to perform clustering.
3. **PCA Implementation:**
   - ❖ Develop a class named PCA to implement the standard PCA method.
   - ❖ Use PCA to reduce the dimensionality of the dataset and visualize the principal components with dimension being 2 and 3, respectively.
4. **Nonlinear Autoencoder Implementation:**
   - ❖ Develop a class named NonlinearAutoEncoder to implement a nonlinear autoencoder.
   - ❖ Use the autoencoder to reduce the dimensionality of the dataset, with dimension being 2 and 3, respectively.

5. **Clustering with Reduced Dimensions:**

   - Apply K-Means++ and soft K-means on the reduced dimensions obtained from PCA and nonlinear autoencoder, respectively.
   - Compare the clustering results with the ones obtained from 1 and 2.

6. **MLP for Multi-Class Classification:**

   - Apply a Multi-Layer Perceptron to solve the multi-class classification problem.
   - Compare the performance of the MLP with the clustering results from 1, 2, and 5 in terms of accuracy. For clustering results, you could add corresponding labels to the clusters.

7. **SVM and SVM with Gaussian Kernel:**

   - Develop algorithms for SVM and SVM with Gaussian kernel using Python and NumPy.

8. **AdaBoost Algorithm:**

   - Develop the AdaBoost algorithm using Python and NumPy.

9. **Binary Classification:**

   - Remove the data with label 2 to create a binary classification dataset.
   - Apply MLP, SVM, SVM with Gaussian, and AdaBoost to solve the binary classification problem.
   - Compare the performance of these methods.

Results:

- Present the results of each method, including clustering results, classification accuracy, and any other relevant metrics.
- Compare the results from the different methods and discuss the implications. Discuss the advantages and disadvantages of each method in the context of the wheat seed dataset.
- Summarize the findings and provide conclusions based on the performance of the different machine learning techniques.